Circuit Design Methodologies for Soft Error Resilience in Nano-Scaled CMOS Technologies

N. George and R.W.Mann
ECE 632 – Fall 2008
University of Virginia
[njg3v,rwm3p]@virginia.edu

ABSTRACT

The aggressive scaling of CMOS technologies continues in an attempt to stay on track with Moore's Law. Methods that lower device sizes, operating voltages, and increase operating frequencies with each technology generation also increase susceptibility of these devices to soft errors. We analyze this trend in both SRAM and logic structures by evaluating the minimum charge required to cause an upset in these circuits. We explore methods to increase Q_{crit} in SRAM using novel topologies that increase node capacitance. We also extend bit interleaving, a well-known method for protecting memories against multi-bit errors, to combinational and sequential logic to render immunity to multi-bit errors.

1. INTRODUCTION

As CMOS technologies continue to scale into the nanometer regime, the potential soft error rate (SER) impact to both SRAM and logic circuits is increasing. The evolution of CMOS technologies has been characterized by aggressive device scaling in order to achieve higher performance with lower power consumption. Although the reduced dimensions and lower operating voltages driven by scaling bring down power consumption, the reduced charge stored on the node or bit increase the susceptibility to radiation induced upset or *soft errors* or single-event upsets (SEUs). Since the initial report by May and Woods that soft errors in dynamic memory circuits could be attributed to radiation (specifically alpha particles) [1], the effect of soft errors caused by radiation on semiconductor devices has become well known.

The two primary sources of SEU inducing radiation are from either terrestrial radiation or from radioactive isotopes within materials used in the integrated circuit fabrication process. Terrestrial or cosmic radiation interaction with the earth's atmosphere results in a shower of neutron particles that exhibit a large range of energies from ~1MeV to several 100MeV. These high energy neutrons can interact with silicon through elastic and inelastic recoil or by spallation where the silicon atom is shattered into heavy and one or several light particles [2]. This process produces a charge cloud of electron-hole pairs that, when in close proximity (<~2μm) to one or more sensitive neighboring circuit nodes, may result in a single or multi-bit error.

The second form of radiation which predominately originates from impurities within the materials used in modern interconnect technology is the alpha particle. The alpha particle is essentially a doubly ionized (He) atom consisting of 2 protons and 2 neutrons. The alpha particle originating from the impurities and isotopes used in the interconnect materials has an initial energy in the range of ~0.6- 7.7 MeV. Because the alpha particle is ionized, it interacts with the silicon lattice to produce a column of electron-hole pairs along the trajectory path, which can cause an upset of the sensitive circuit node.

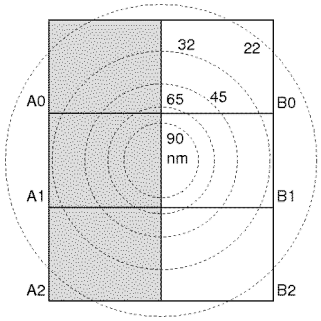

Figure 1 Area covered by a 2μm radius with respect to the area of two 3-bit registers at various technology nodes

An additional consequence of scaling is the probability of multi-bit errors, which occur when single or multiple particle strike induces errors on multiple bits or logic nodes. Multi-bit errors can be classified as spatial and temporal multi-bit errors [3]. If a single particle strike upsets more than one bit in the neighborhood of the struck bit, it is referred to as a spatial multi-bit error. If two or more independent particle strikes (possibly at different times) affect multiple bits in the same memory word, it is referred to as a temporal multi-bit error. The probability of a temporal multi-bit error at the 130nm node (where the common industry bit cell size >2μm^2) is sufficiently low (~10^{-7}) that we can ignore its effects [4].

However, as we scale into the nanometer generation of device sizing, more and more bits in a memory array fall under the footprint of a single particle strike, spatial multi-bit errors become a more significant concern, because charge upsets can occur within one to two microns of a junction where the particle strikes[5][6][7] . The number of bit cells within a 2μm radius in a dense SRAM array at the 65nm node is ~25 and will roughly double with every successive node. Figure 1 shows the size of a 2μm radius with respect to typical sequential logic at various technology nodes.

In this paper, we analyze the critical charge (Q_{crit}) for dense SRAM and combinational logic as a function of technology scaling from 130nm to 16nm. We believe this is the first paper to have performed Qcrit investigations extending beyond the 45nm generation. Multi-bit soft error upset in SRAM has been addressed through interleaving which enables errors to be addressed by ECC as single bit faults. The authors explored if a similar method could be applied to combinational and sequential logic. We present the utility and effectiveness of this technique, and an analysis of its overhead.

2. EXPERIMENTAL SETUP & RESULTS

2.1 Technology Scaling and SRAM Q_{crit}

SRAM is commonly used to establish a baseline for Q_{crit} and SER within a technology for three primary reasons, (1) the wide spread use of SRAM in the industry and (2) the use of aggressive device widths in the bit-cell means it will exhibit a relatively high sensitivity to SEU (3) it provides a means of obtaining spatially diagnosable fail information. Because our intent is to explore the effect of scaling on both the SRAM and logic SER, we also begin with the SRAM to establish the fundamental parameters such as current pulse shape and duration for modeling the Q_{crit} across the range of technologies of interest.

The SER for SRAM as well as logic is exponentially dependent on Q_{crit} and proportional to the sensitive node area as shown in Equation 1:

$$SER \propto F \times A_{diff} \times \exp\left(-\frac{Q_{crit}}{Q_s}\right) \qquad \text{Eq. 1}$$

where F is the particle flux, A_{diff} is the critical or sensitive charge collection area, Q_{crit} is the critical amount of charge required to flip the bit and Q_s is the charge collection efficiency. Cell designs topologies which minimize A_{diff} and increase Q_{crit} are therefore preferred. The charge collection efficiency Q_s is modulated by factors such as NWELL and PWELL depth, use of retrograde well doping profiles and use of triple well are established by technology developers. The amount of charge collected is typically much less than the total charge generated.

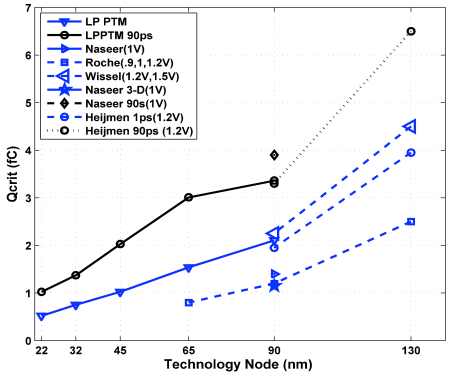

Figure 2 Comparison of published literature values for critical charge and the values for the LP PTM technology.

To evaluate the critical charge required to flip the SRAM cell, we used the double exponential pulse with 0.1ps rise and 5.5ps fall to simulate the alpha strike and 90ps fall for charge collection response associated with a neutron collision charge cloud [8][9]. Figure 2 shows the modeled Q_{crit} based on the LP PTMs [10] compared to values published by others. The simulations were performed with Ocean scripts to perform a binary search method to identify the Q_{crit} using the Spectre simulation environment. The overall slope of the trends from 130 to 90 is roughly 2x steeper than our trends extending to 22nm. This is attributed to reduced voltage scaling for sub 100nm technologies. The availability of several PTM variations allowed us to compare the Q_{crit} trends across a range of high performance (HP) and low power (LP) technologies shown in figure 3. The trends show that Q_{crit} is decreasing with each generation with a slope of approximately 0.023fC/nm for the alpha simulation and 0.046fC/nm for simulation of the neutron charge collection response.

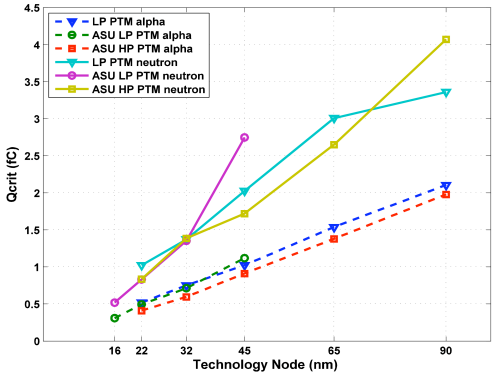

Figure 3 Trend in Q_{crit} for HP and LP technologies down to 16nm. Both alpha and neutron charge collection results are shown.

SRAM bit cell design topologies do not significantly change the overall reduction in Q_{crit} associated with scaling. While factors in the cell design such as node capacitance and node-to-node capacitance can impact Q_{crit}, we show that the amount of capacitance required to provide significant deviation from the scaling trends are somewhat large and would not be without some area or penalty in additional processing and complexity cost. The effect of both node capacitance and node-node capacitance on Q_{crit} for scaled high performance technologies is shown in Figure 4.

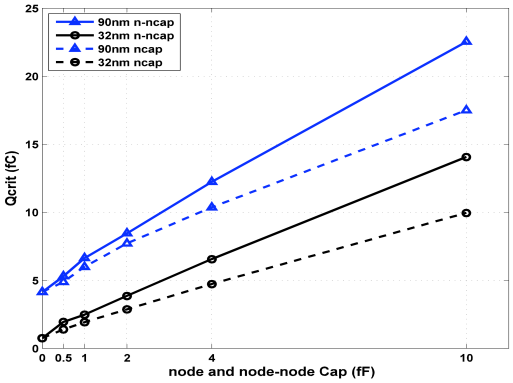

Figure 4 The effect of capacitance adders to the Q_{crit}. Node to node capacitance (solid line) is more beneficial.

From figure 4 it is observed that the Q_{crit} is more significantly improved by the addition of node-node capacitance than the straight forward addition of node capacitance. Because the sensitive node area (A_{diff}) is shrinking by roughly a factor of 2 for every generation, and the collection efficiency is decreasing, the net SER per bit is expected to remain roughly constant with continued scaling.

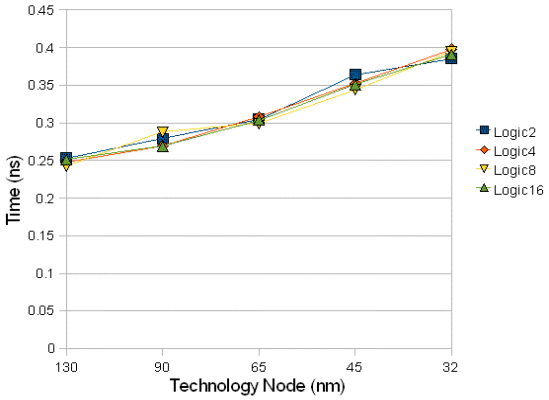

Figure 5 Width of vulnerable latching window around latching edge for various logic depths

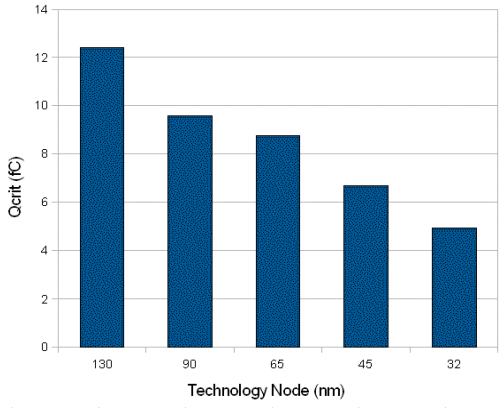

Figure 6 Critical charge of a positive register at various technology nodes

2.2 Characterization of Logic Q_{crit}

To characterize critical charge in combinational logic, we used a simplified model of a pipeline stage as a chain of NAND gates of variable length between two sequential elements. A transient particle striking a node in combinational logic manifests as a voltage pulse that travels through the chain of logic. If this voltage glitch appears at the input of the sequential element during a time interval corresponding to the setup plus hold time of the latching edge, an erroneous value could get latched. This model was used to determine the window around the latching edge during which a transient pulse appearing well ahead in the logic chain can get latched at the output sequential element.

We measured the vulnerable latching window for logic lengths corresponding to 2, 4, 8 and 16 NAND chains with positive registers as input and output storage elements. Each of these cases was simulated using PTMs ranging from 90-35nm. A simulated current pulse similar to a neutron strike was injected into the node corresponding to 2(4, 8, or 16) gates from the output sequential element. An Ocean script was used to determine the setup and hold times accurate up to 10ps. Figure 5 shows that the width of this vulnerable window increases as devices are scaled across each generation. This means that the vulnerability of logic to soft errors is worsening at a fast rate because of two reasons – increasing operating frequencies and widening latching windows.

The mid point of this window was chosen to determine critical charge. This point in time was chosen because of the definition of Q_{crit}. The mid point of the window would correspond to the most vulnerable time of latching a wrong value and the smallest current pulse that upsets the circuit would be its true critical charge. A separate Ocean script was used to find the smallest current pulse that caused an upset in the output. Binary search algorithms were used for both (window and Q_{crit} determination) scripts to minimize simulation times. The results of this simulation are summarized in Figure 6. It can be seen that the critical charge is smaller for smaller devices and that logic Q_{crit} numbers are soon approaching that of SRAM.

These results indicate the increasing vulnerability of logic to soft errors and arguments from Figure 1 indicate the vulnerability of sequential logic to multi-bit upsets and provides motivation for integrating fault protection not only into SRAM but also logic circuits. We made use of single error detection methods available for logic together with interleaving in layout to provide immunity to multiple bit upsets.

2.3 Self Checking Adders

Parity prediction is a scheme that has been proposed by [11] for implementation on arithmetic and logical operations for error detection based on arithmetic codes on inputs and outputs. It is based on the fact that, in an adder, the parity of the result can be computed using the parity of the inputs and the parity of the carry chain using the relation $P_S = P_A \oplus P_B \oplus P_C$. Where, P_S is the parity of the output (sum) vector, P_A and P_B are the parities of the input (operands) vector, and P_C is the parity of the carry chain. Thus we have the parity of the result being computed from two places. One, the actual result, and the other using the *prediction* mechanism derived from the expression above.

This technique requires a redundant carry chain to ensure correctness. A formal proof of this requirement is available in [11]. This design is fault secure with respect to single faults in either the inputs or in the logic circuit. This adder circuit will be used as the baseline for comparing the proposed method to extend the fault-secure property to include coverage of single faults causing multiple errors.

2.4 Logic Interleaving

To explore the utility and to analyze overheads of interleaving logic, we laid out a 4-bit Brent-Kung adder protected using the carry-checking/parity prediction methodology presented in [11] using a 90nm process. This design provides fault protection to the combinational-sequential logic combination (input and output registers, logic for sum/carry generation and logic for fault protection) up to single faults. This layout was modified by spacing out nodes sufficiently to avoid worst-case charge collection. Further spacing was required to provide enough spacing to enable interleaving. The layout was DRC and LVS verified. Extracted parasitics were then simulated to determine delays along various paths in the circuit.

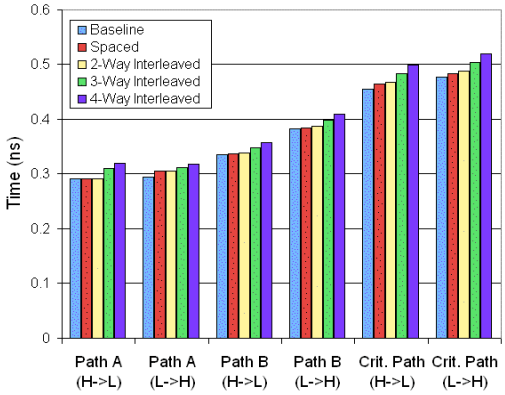

Figure 7 Absolute delays and overheads on various paths

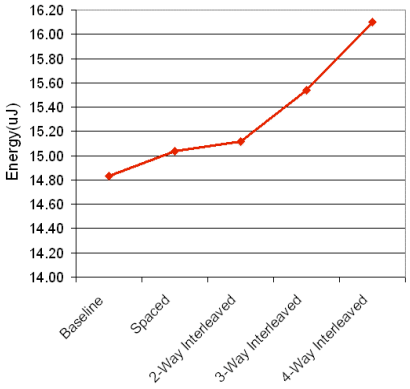

Figure 8 Total energy consumed for a unit of work

A closer examination of Figure 1 reveals that interleaving 2 registers is sufficient when using a 90nm process to prevent multiple upsets in a given byte/word. 2-way interleaving is however not sufficient at the next three generations namely 65, 45 and 32nm. 3-way interleaving would be required at these nodes to guarantee that worst case charge collection would not affect multiple bits from the same word. Similarly 3-way interleaving would not suffice at the 22nm node and would require 4-way interleaving. To get a comparative estimate of the delay overheads of 3- and 4-way interleaving, these were implemented at the 90nm process. Delays of 2-way, 3-way and 4-way interleaving were determined and are shown in Figure 7 as a comparison to the spaced and baseline designs.

It is worth noting that the delay overhead of 2-way interleaving compared to the design with spaced out nodes, is less than 1% for the critical path. The biggest advantage of incurring this overhead is that the area overhead drops to 0 (spaced designs do not have sufficient area to fit another bit-slice). The delay overhead of the 4-way interleaved design is 8.25% compared to the baseline for the critical path in a 90nm process. Figure 8 shows the total energy consumed by each of the designs to complete the same amount of work. It can be said that the amount of total energy goes up almost linearly with the degree of interleaving.

3. DISCUSSION & CONCLUSIONS

To the best of our knowledge, this is the first paper that explores the effects of scaling on Qcrit beyond the 45nm node. We find that the value of Q_{crit} for SRAM will continue to be reduced by roughly 0.023fC/nm with continued scaling. This slope was consistent for high performance as well as low power PTM technologies. Although the change in Q_{crit} with scaling has slowed with voltage scaling, we find that technology levers to effect significant improvement are limited. We examined device width, node capacitance and node-node capacitance and conclude that the most promising technology option is node-node capacitance but the cost in either density and/or process complexity is likely prohibitive. Circuit solutions such as interleaving and ECC will therefore be the most promising path for designs that require increased soft error immunity. We also found that the critical charge of logic circuits is soon approaching that of SRAM. The problem is worsened by a number of factors such as increasing operating frequencies, widening latching windows, and increasing number of sensitive nodes within a susceptible region. That provided motivation for exploring interleaving logic as a method to tolerate SEUs. We found that it could be an effective method to overcome the problem of multi-bit errors because is comes with negligible timing overheads (~1%) and almost linear energy overheads.

4. REFERENCES

[1] T. May, M. Woods, "A New Physical Mechanism for Soft Error in Dynamic Memories." *Proc. 1978 Int'l Rel. Physics Symposium*, pp.33-40, 1978

[2] Lambert, D.; Baggio, J.; Ferlet-Cavrois, V.; Flament, O.; Saigne, F.; Sagnes, B.; Buard, N.; Carriere, T., "Neutron-induced SEU in bulk SRAMs in terrestrial environment: Simulations and experiments," *Nuclear Science, IEEE Transactions on* , Vol.51, no.6, pp. 3435-3441, Dec. 2004

[3] P. Shivakumar, M. Kistler, S. Keckler, D. Berger, and L. Alvisi, "Modeling the Effect of Technology Trends on the Soft Error Rate of Combinational Logic", *International Conference of Dependable Systems and Networks*, June 2002

[4] Dodd, P.E.; Massengill, L.W., "Basic mechanisms and modeling of single-event upset in digital microelectronics," *Nuclear Science, IEEE Transactions on*, vol.50, no.3, pp. 583-602, June 2003

[5] N. Seifert; P. Slankard; M. Kirsch; B. Narasimham; V. Zia; C. Brookreson; A. Vo; S. Mitra; B. Gill; J. Maiz, "Radiation-Induced Soft Error Rates of Advanced CMOS Bulk Devices," *Reliability Physics Symposium Proceedings, 2006. 44th Annual., IEEE International* , vol., no., pp.217-225, March 2006

[6] R. Baumann, "The Impact of Technology Scaling on Soft Error Rate Performance and Limits to the Efficacy of Error Correction," in *Proc. IEEE Inf. Dw. Meet. (IEDM),* pp. 329-332,2002

[7] Amusan, O. A.; Witulski, A. F.; Massengill, L. W.; Bhuva, B. L.; Fleming, P. R.; Alles, M. L.; Sternberg, A. L.; Black, J. D.; Schrimpf, R. D., "Charge Collection and Charge Sharing in a 130 nm CMOS Technology," *Nuclear Science, IEEE Transactions on* , vol.53, no.6, pp.3253-3258, Dec. 2006

[8] Naseer, R.; Boulghassoul, Y.; Draper, J.; DasGupta, S.; Witulski, A., "Critical Charge Characterization for Soft Error Rate Modeling in 90nm SRAM," *Circuits and Systems, 2007. ISCAS 2007. IEEE International Symposium on* , vol., no., pp.1879-1882, 27-30 May 2007

[9] T. Heijmen, D. Giot, P. Roche, "Factors That Impact the Critical Charge of Memory Elements," *12th IEEE International On-Line Testing Symposium (IOLTS'06)*, pp.57-62, 2006

[10] W. Zhao and Y. Cao, "New generation of predictive technology model for sub-45 nm early design exploration," *Electron Devices, IEEE Transactions on*, vol. 53, no. 11, pp. 2816–2823, 2006

[11] M. Nicolaidis, "Carry Checking/Parity Prediction Adders and ALUs," *IEEE Transactions on VLSI Systems*, Vol. 11, No. 1, February 2003